

A P P L I C A T I O N

for

UNITED STATES LETTERS PATENT

on

A METHOD OF INCREASING COMPLEMENTARITY

IN A HETERODUPLEX

by

HAL S. PADGETT

JOHN A. LINDBO

WAYNE P. FITZMAURICE

CERTIFICATE OF MAILING BY "EXPRESS MAIL"

"EXPRESS MAIL" MAILING LABEL NUMBER EL 857042906US

DATE OF DEPOSIT February 1, 2002

I HEREBY CERTIFY THAT THIS PAPER OR FEE IS BEING DEPOSITED
WITH THE UNITED STATES POSTAL SERVICE "EXPRESS MAIL POST
OFFICE TO ADDRESSEE" SERVICE UNDER 37 C.F.R. 1.10 ON THE
DATE INDICATED ABOVE AND IS ADDRESSED TO: THE
COMMISSIONER FOR PATENTS, WASHINGTON, D.C. 20231.

James J. Wong
(TYPED OR PRINTED NAME OR PERSON MAILING PAPER OR FEE)

James J. Wong
(SIGNATURE OF PERSON MAILING PAPER OR FEE)

Sheets of Drawings: Two (2)

Docket No.: P-LG 4878

Attorneys
CAMPBELL & FLORES LLP
4370 La Jolla Village Drive, 7th Floor
San Diego, California 92122
USPTO CUSTOMER NO. 23601

**A METHOD OF INCREASING
COMPLEMENTARITY IN A HETERODUPLEX**

This application claims the benefit of U.S. 5 Provisional Application No. 60/266,386, filed February 2, 2001, and of U.S. Provisional Application No. 60/268,785, Filed February 14, 2001, and are both incorporated herein by reference.

10

BACKGROUND OF THE INVENTION

FIELD OF THE INVENTION

The invention relates generally to molecular 15 biology and more specifically to methods of generating populations of related nucleic acid molecules.

BACKGROUND INFORMATION

20 DNA shuffling is a powerful tool for obtaining recombinants between two or more DNA sequences to evolve them in an accelerated manner. The parental, or input, DNAs for the process of DNA shuffling are typically mutants or variants of a given gene that have some 25 improved character over the wild-type. The products of DNA shuffling represent a pool of essentially random reassortments of gene sequences from the parental DNAs that can then be analyzed for additive or synergistic effects resulting from new sequence combinations.

Recursive sequence reassortment is analogous to an evolutionary process where only variants with suitable properties are allowed to contribute their genetic material to the production of the next generation.

5 Optimized variants are generated through DNA shuffling-mediated sequence reassortment followed by testing for incremental improvements in performance. Additional cycles of reassortment and testing lead to the generation of genes that contain new combinations of the genetic

10 improvements identified in previous rounds of the process. Reassorting and combining beneficial genetic changes allows an optimized sequence to arise without having to individually generate and screen all possible sequence combinations.

15

This differs sharply from random mutagenesis, where subsequent improvements to an already improved sequence result largely from serendipity. For example, in order to obtain a protein that has a desired set of

20 enhanced properties, it may be necessary to identify a mutant that contains a combination of various beneficial mutations. If no process is available for combining these beneficial genetic changes, further random mutagenesis will be required. However, random

25 mutagenesis requires repeated cycles of generating and screening large numbers of mutants, resulting in a process that is tedious and highly labor intensive. Moreover, the rate at which sequences incur mutations with undesirable effects increases with the information

30 content of a sequence. Hence, as the information content, library size, and mutagenesis rate increase, the

ratio of deleterious mutations to beneficial mutations will increase, increasingly masking the selection of further improvements. Lastly, some computer simulations have suggested that point mutagenesis alone may often be 5 too gradual to allow the large-scale block changes that are required for continued and dramatic sequence evolution.

There are a number of different techniques used 10 for random mutagenesis. For example, one method utilizes error-prone polymerase chain reaction (PCR) for creating mutant genes in a library format, (Cadwell and Joyce, 1992; Gram et al., 1992). Another method is cassette mutagenesis (Arkin and Youvan, 1992; Delagrange et al., 15 1993; Delagrange and Youvan, 1993; Goldman and Youvan, 1992; Hermes et al., 1990; Oliphant et al., 1986; Stemmer et al., 1993) in which the specific region to be optimized is replaced with a synthetically mutagenized oligonucleotide.

20

Error-prone PCR uses low-fidelity polymerization conditions to introduce a low level of point mutations randomly over a sequence. A limitation to this method, however, is that published error-prone 25 PCR protocols suffer from a low processivity of the polymerase, making this approach inefficient at producing random mutagenesis in an average-sized gene.

In oligonucleotide-directed random mutagenesis, 30 a short sequence is replaced with a synthetically mutagenized oligonucleotide. To generate combinations of

distant mutations, different sites must be addressed simultaneously by different oligonucleotides. The limited library size that is obtained in this way, relative to the library size required to saturate all 5 sites, means that many rounds of selection are required for optimization. Mutagenesis with synthetic oligonucleotides requires sequencing of individual clones after each selection round followed by grouping them into families, arbitrarily choosing a single family, and 10 reducing it to a consensus motif. Such a motif is resynthesized and reinserted into a single gene followed by additional selection. This step creates a statistical bottleneck, is labor intensive, and is not practical for many rounds of mutagenesis.

15

For these reasons, error-prone PCR and oligonucleotide-directed mutagenesis can be used for mutagenesis protocols that require relatively few cycles of sequence alteration, such as for sequence fine-tuning, 20 but are limited in their usefulness for procedures requiring numerous mutagenesis and selection cycles, especially on large gene sequences.

As discussed above, prior methods for producing 25 improved gene products from randomly mutated genes are of limited utility. One recognized method for producing a wide variety of randomly reassorted gene sequences uses enzymes to cleave a long nucleotide chain into shorter pieces. The cleaving agents are then separated from the 30 genetic material, and the material is amplified in such a manner that the genetic material is allowed to reassemble

as chains of polynucleotides, where their reassembly is either random or according to a specific order.

((Stemmer, 1994a; Stemmer, 1994b), US Patent No.

5,605,793, US Patent No. 5,811,238, US Patent No.

5 5,830,721, US Patent No. 5,928,905, US Patent No.

6,096,548, US Patent No. 6,117,679, US Patent No.

6,165,793, US Patent No. 6,153,410). A variation of this method uses primers and limited polymerase extensions to generate the fragments prior to reassembly (US Patent No.

10 5,965,408, US Patent No. 6,159,687).

However, both methods have limitations. These methods suffer from being technically complex. This limits the applicability of these methods to facilities 15 that have sufficiently experienced staffs. In addition there are complications that arise from the reassembly of molecules from fragments, including unintended mutagenesis and the increasing difficulty of the reassembly of large target molecules of increasing size, 20 which limits the utility of these methods for reassembling long polynucleotide strands.

Another limitation of these methods of fragmentation and reassembly-based gene shuffling is 25 encountered when the parental template polynucleotides are increasingly heterogeneous. In the annealing step of those processes, the small polynucleotide fragments depend upon stabilizing forces that result from base-pairing interactions to anneal properly. As the small 30 regions of annealing have limited stabilizing forces due to their short length, annealing of highly complementary

sequences is favored over more divergent sequences. In such instances these methods have a strong tendency to regenerate the parental template polynucleotides due to annealing of complementary single-strands from a 5 particular parental template. Therefore, the parental templates essentially reassemble themselves creating a background of unchanged polynucleotides in the library that increases the difficulty of detecting recombinant molecules. This problem becomes increasingly severe as 10 the parental templates become more heterogeneous, that is, as the percentage of sequence identity between the parental templates decreases. This outcome was demonstrated by Kikuchi, *et al.*, (Gene 243:133-137, 2000) who attempted to generate recombinants between *xyle* and 15 *nahH* using the methods of family shuffling reported by Patten *et al.*, 1997; Crameri *et al.*, 1998; Harayama, 1998; Kumamaru *et al.*, 1998; Chang *et al.*, 1999; Hansson *et al.*, 1999). Kikuchi, *et al.*, found that essentially no recombinants (<1%) were generated. They also 20 disclosed a method to improve the formation of chimeric genes by fragmentation and reassembly of single-stranded DNAs. Using this method, they obtained chimeric genes at a rate of 14 percent, with the other 86 percent being parental sequences.

25

The characteristic of low-efficiency recovery of recombinants limits the utility of these methods for generating novel polynucleotides from parental templates with a lower percentage of sequence identity, that is, 30 parental templates that are more diverse. Accordingly,

there is a need for a method of generating gene sequences that addresses these needs.

The present invention provides a method that
5 satisfies the aforementioned needs, and also provides
related advantages as well.

Summary of the Invention

10 The present invention provides a method for reassorting mutations among related polynucleotides, *in vitro*, by forming heteroduplex molecules and then addressing the mismatches such that sequence information at sites of mismatch is transferred from one strand to
15 the other. In one preferred embodiment, the mismatches are addressed by incubating the heteroduplex molecules in a reaction containing a mismatch nicking enzyme, a polymerase with a 3' to 5' proofreading activity in the presence of dNTPs, and a ligase. These respective
20 activities act in concert such that, at a given site of mismatch, the heteroduplex is nicked, unpaired bases are excised then replaced using the opposite strand as a template, and nicks are sealed. Output polynucleotides are amplified before cloning, or cloned directly and
25 tested for improved properties. Additional cycles of mismatch resolution reassortment and testing lead to further improvement.

PROOFREADING ENZYME

Brief Description of the Figures

Figure 1 depicts the process of Genetic ReAssortment by Mismatch Resolution (GRAMMR). Reassortment is contemplated between two hypothetical polynucleotides differing at at least two nucleotide positions. Annealing between the top strand of A and the bottom strand of B is shown which results in mismatches at the two positions. After the process of reassortment mismatch resolution, four distinct product polynucleotides are seen, the parental types A and B, and the reassorted products X and Y.

Figure 2 depicts an exemplary partially complementary nucleic acid population of two molecules. Figure 2A shows the sequence of two nucleic acid molecules "X" and "Y" having completely complementary top/bottom strands 1+/2- and 3+/4-, respectively. The positions of differing nucleotides between the nucleic acids X and Y are indicated (*). Figure 2B shows possible combinations of single strands derived from nucleic acids X and Y after denaturing and annealing and indicates which of those combinations would comprise a partially complementary nucleic acid population of two.

Definitions

As used herein the term "amplification" refers to a process where the number of copies of a polynucleotide is increased.

As used herein, "annealing" refers to the formation of at least partially double stranded nucleic acid by hybridization of at least partially complementary nucleotide sequences. A partially double stranded nucleic acid can be due to the hybridization of a smaller nucleic acid strand to a longer nucleic acid strand, where the smaller nucleic acid is 100% identical to a portion of the larger nucleic acid. A partially double stranded nucleic acid can also be due to the hybridization of two nucleic acid strands that do not share 100% identity but have sufficient homology to hybridize under a particular set of hybridization conditions.

15

As used herein, "clamp" refers to a unique nucleotide sequence added to one end of a polynucleotide, such as by incorporation of the clamp sequence into a PCR primer. The clamp sequences are intended to allow amplification only of polynucleotides that arise from hybridization of strands from different parents (*i.e.*, heteroduplex molecules) thereby ensuring the production of full-length hybrid products as described previously (Skarfstad, *J. Bact.*, vol 182, No 11, P. 3008-3016).

25

As used herein the term "cleaving" means digesting the polynucleotide with enzymes or otherwise breaking phosphodiester bonds within the polynucleotide.

As used herein the term "complementary basepair" refers to the correspondence of DNA (or RNA) bases in the double helix such that adenine in one strand 5 is opposite thymine (or uracil) in the other strand and cytosine in one strand is opposite guanine in the other.

As used herein the term "complementary to" is used herein to mean that the complementary sequence is 10 identical to the reverse-complement of all or a portion of a reference polynucleotide sequence or that each nucleotide in one strand is able to form a base-pair with a nucleotide, or analog thereof in the opposite strand. For illustration, the nucleotide sequence "TATAc" is 15 complementary to a reference sequence "GTATA".

As used herein, "denaturing" or "denatured," when used in reference to nucleic acids, refers to the conversion of a double stranded nucleic acid to a single 20 stranded nucleic acid. Methods of denaturing double stranded nucleic acids are well known to those skilled in the art, and include, for example, addition of agents that destabilize base-pairing, increasing temperature, decreasing salt, or combinations thereof. These factors 25 are applied according to the complementarity of the strands, that is, whether the strands are 100% complementary or have one or more non-complementary nucleotides.

As used herein the term "desired functional property" means a phenotypic property, which include but are not limited to, encoding a polypeptide, promoting 5 transcription of linked polynucleotides, binding a protein, improving the function of a viral vector, and the like, which can be selected or screened for. Polynucleotides with such desired functional properties, can be used in a number of ways, which include but are 10 not limited to expression from a suitable plant, animal, fungal, yeast, or bacterial expression vector, integration to form a transgenic plant, animal or microorganism, expression of a ribozyme, and the like.

15 As used herein the term "DNA shuffling" is used herein to indicate recombination between substantially homologous but non-identical sequences.

20 As used herein, the term "effective amount" refers to the amount of an agent necessary for the agent to provide its desired activity. For the present invention, this determination is well within the knowledge of those of ordinary skill in the art.

25 As used herein the term "exonuclease" refers to an enzyme that cleaves nucleotides one at a time from an end of a polynucleotide chain, that is, an enzyme that hydrolyzes phosphodiester bonds from either the 3' or 5' terminus of a polynucleotide molecule. Such 30 exonucleases, include but are not limited to T4 DNA polymerase, T7 DNA polymerase, *E. coli* Pol 1, and Pfu DNA

polymerase. The term "exonuclease activity" refers to the activity associated with an exonuclease. An exonuclease that hydrolyzes in a 3' to 5' direction is said to have "3' to 5' exonuclease activity." Similarly 5 an exonuclease with 5' to 3' activity is said to have "5' to 3' exonuclease activity." It is noted that some exonucleases are known to have both 3' to 5', 5' to 3' activity, such as, *E.coli* Pol I.

10 As used herein, "Genetic Reassortment by Mismatch Resolution (GRAMMR)" refers to a method for reassorting sequence variations among related polynucleotides by forming heteroduplex molecules and then addressing the mismatches such that information is 15 transferred from one strand to the other.

As used herein, "granularity" refers to the amount of a nucleic acid's sequence information that is transferred as a contiguous sequence from a template 20 polynucleotide strand to a second polynucleotide strand. As used herein, "template sequence" refers to a first single stranded polynucleotide sequence that is partially complementary to a second polynucleotide sequence such that treatment by GRAMMR results in transfer of genetic 25 information from the template strand to the second strand.

The larger the units of sequence information transferred from a template strand, the higher the 30 granularity. The smaller the blocks of sequence information transferred from the template strand, the

lower or finer the granularity. Lower granularity indicates that a DNA shuffling or reassortment method is able to transfer smaller discrete blocks of genetic information from the template strand to the second strand. The advantage of a DNA shuffling or reassortment method with lower granularity is that it is able to resolve smaller nucleic acid sequences from others, and to transfer the sequence information. DNA shuffling or reassortment methods that return primarily high granularity are not readily able to resolve smaller nucleic acid sequences from others.

As used herein the term "heteroduplex polynucleotide" refers to a double helix polynucleotide formed by annealing single strands, typically separate strands, where the strands are non-identical. A heteroduplex polynucleotide may have unpaired regions existing as single strand loops or bubbles. A heteroduplex polynucleotide region can also be formed by one single-strand polynucleotide wherein partial self-complementarity allows the formation of a stem-loop structure where the annealing portion of the strand is non-identical.

As used herein the term "heteroduplex DNA" refers to a DNA double helix formed by annealing single strands, typically separate strands), where the strands are non-identical. A heteroduplex DNA may have unpaired regions existing as single strand loops or bubbles. A heteroduplex DNA region can also be formed by one single-strand polynucleotide wherein partial self-

complementarity allows the formation of a stem-loop structure where the annealing portion of the strand is non-identical.

5 As used herein the term "homologous" means that one single-stranded nucleic acid sequence may hybridize to an at least partially complementary single-stranded nucleic acid sequence. The degree of hybridization may depend on a number of factors including the amount of
10 identity between the sequences and the hybridization conditions such as temperature and salt concentrations as discussed later.

15 Nucleic acids are "homologous" when they are derived, naturally or artificially, from a common ancestor sequence. During natural evolution, this occurs when two or more descendent sequences diverge from a parent sequence over time, *i.e.*, due to mutation and natural selection. Under artificial conditions,
20 divergence occurs, *e.g.*, in one of two basic ways. First, a given sequence can be artificially recombined with another sequence, as occurs, *e.g.*, during typical cloning, to produce a descendent nucleic acid, or a given sequence can be chemically modified, or otherwise
25 manipulated to modify the resulting molecule. Alternatively, a nucleic acid can be synthesized *de novo*, by synthesizing a nucleic acid that varies in sequence from a selected parental nucleic acid sequence. When
30 there is no explicit knowledge about the ancestry of two nucleic acids, homology is typically inferred by sequence comparison between two sequences. Where two nucleic acid

sequences show sequence similarity over a significant portion of each of the nucleic acids, it is inferred that the two nucleic acids share a common ancestor. The precise level of sequence similarity that establishes 5 homology varies in the art depending on a variety of factors.

For purposes of this disclosure, two nucleic acids are considered homologous where they share 10 sufficient sequence identity to allow GRAMMR-mediated information transfer to occur between the two nucleic acid molecules.

As used herein the term "identical" or 15 "identity" means that two nucleic acid sequences have the same sequence or a complementary sequence. Thus, "areas of identity" means that regions or areas of a polynucleotide or the overall polynucleotide are identical or complementary to areas of another 20 polynucleotide.

As used herein the term "increase in percent complementarity" means that the percentage of complementary base-pairs in a heteroduplex molecule is 25 made larger.

As used herein the term, "ligase" refers to an enzyme that rejoins a broken phosphodiester bond in a nucleic acid.

As used herein the term "mismatch" refers to a base-pair that is unable to form normal base-pairing interactions (i.e., other than "A" with "T" (or "U"), or "G" with "C").

5

As used herein the term "mismatch resolution" refers to the conversion of a mismatched base-pair into a complementary base-pair.

10

As used herein the term "mutations" means changes in the sequence of a wild-type or reference nucleic acid sequence or changes in the sequence of a polypeptide. Such mutations can be point mutations such as transitions or transversions. The mutations can be 15 deletions, insertions or duplications.

20

As used herein the term "nick translation" refers to the property of a polymerase where the combination of a 5'-to-3' exonuclease activity with a 5'-to-3' polymerase activity allows the location of a single-strand break in a double-stranded polynucleotide (a "nick") to move in the 5'-to-3' direction.

25

As used herein, the term "nucleic acid" or "nucleic acid molecule" means a polynucleotide such as deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) and encompasses single-stranded and double-stranded nucleic acid as well as an oligonucleotide. Nucleic acids useful in the invention include genomic DNA, cDNA, mRNA and 30 synthetic oligonucleotides, and can represent the sense strand, the anti-sense strand, or both. A nucleic acid

generally incorporates the four naturally occurring nucleotides adenine, guanine, cytosine, and thymidine/uridine. An invention nucleic acid can also incorporate other naturally occurring or non-naturally occurring nucleotides, including derivatives thereof, so long as the nucleotide derivatives can be incorporated into a polynucleotide by a polymerase at an efficiency sufficient to generate a desired polynucleotide product.

10 As used herein, a "parental nucleic acid" refers to a double stranded nucleic acid having a sequence that is 100% identical to an original single stranded nucleic acid in a starting population of partially complementary nucleic acids. Parental nucleic acids would include, for example in the illustration of Figure 2, nucleic acids X and Y if partially complementary nucleic acid combinations 1+/4- or 2-/3+ were used as a starting population in an invention method.

20 As used herein, "partially complementary" refers to a nucleic acid having a substantially complementary sequence to another nucleic acid but that differs from the other nucleic acid by at least two or 25 more nucleotides. As used herein, "partially complementary nucleic acid population" refers to a population of nucleic acids comprising nucleic acids having substantially complementary sequences but no nucleic acids having an exact complementary sequence for 30 any other member of the population. As used herein, any member of a partially complementary nucleic acid

population differs from another nucleic acid of the population, or the complement thereto, by two or more nucleotides. As such, a partially complementary nucleic acid specifically excludes a population containing 5 sequences that are exactly complementary, that is, a complementary sequence that has 100% complementarity. Therefore, each member of such a partially complementary nucleic acid population differs from other members of the population by two or more nucleotides, including both 10 strands. One strand is designated the top strand, and its complement is designated the bottom strand. As used herein, "top" strand refers to a polynucleotide read in the 5' to 3' direction and the "bottom" its complement. It is understood that, while a sequence is referred to as 15 bottom or top strand, such a designation is intended to distinguish complementary strands since, in solution, there is no orientation that fixes a strand as a top or bottom strand.

20 For example, a population containing two nucleic acid members can be derived from two double stranded nucleic acids, with a potential of using any of the four strands to generate a single stranded partially complementary nucleic acid population. An example of 25 potential combinations of strands of two nucleic acids that can be used to obtain a partially complementary nucleic acid population of the invention is shown in Figure 2. The two nucleic acid sequences that are potential members of a partially complementary nucleic 30 acid population are designated "X" (AGATCAATTG; SEQ ID NO:1) and "Y" (AGACCGATTG; SEQ ID NO:2) (Figure 2A). The

nucleic acid sequences differ at two positions (positions 4 and 6 indicated by "*"). The "top" strand of nucleic acids X and Y are designated "1+" and "3+," respectively, and the "bottom" strand of nucleic acids X and Y are 5 designated "2-" and "4-," respectively.

Figure 2B shows the possible combinations of the four nucleic acid strands. Of the six possible strand combinations, only the combination of 1+/2-, 10 1+/4-, 2-/3+, or 3+/4- comprise the required top and bottom strand of a partially complementary nucleic acid population. Of these top/bottom sequence combinations, only 1+/4- or 2-/3+ comprise an example of a partially complementary nucleic acid population of two different 15 molecules because only these combinations have complementary sequences that differ by at least one nucleotide. The remaining combinations, 1+/2- and 2+/4-, contain exactly complementary sequences and therefore do not comprise a partially complementary nucleic acid 20 population of the invention.

In the above described example of a population of two different molecules, a partially complementary population of nucleic acid molecules excluded 25 combinations of strands that differ by one or more nucleotides but which are the same sense, for example, 1+/3+ or 2-/4-. However, it is understood that such a combination of same stranded nucleic acids can be included in a larger population, so long as the 30 population contains at least one bottom strand and at least one top strand. For example, if a third nucleic

acid "Z," with strands 5+ and 6- is included, the combinations 1+/3+/6- or 2-/4-/5+ would comprise a partially complementary nucleic acid population. Similarly, any number of nucleic acids and their 5 corresponding top and bottom strands can be combined to generate a partially complementary nucleic acid population of the invention so long as the population contains at least one top strand and at least one bottom strand and so long as the population contains no members 10 that are the exact complement.

The populations of nucleic acids of the invention can be about 3 or more, about 4 or more, about 5 or more, about 6 or more, about 7 or more, about 8 or 15 more, about 9 or more, about 10 or more, about 12 or more, about 15 or more, about 20 or more, about 25 or more about 30 or more, about 40 or more, about 50 or more, about 75 or more, about 100 or more, about 150 or more, about 200 or more, about 250 or more, about 300 or 20 more, about 350 or more, about 400 or more, about 450 or more, about 500 or more, or even about 1000 or more different nucleic acid molecules. A population can also contain about 2000 or more, about 5000 or more, about 1×10^4 or more, about 1×10^5 or more, about 1×10^6 or more, 25 about 1×10^7 or more, or even about 1×10^8 or more different nucleic acids. One skilled in the art can readily determine a desirable population to include in invention methods depending on the nature of the desired reassortment experiment outcome and the available 30 screening methods, as disclosed herein.

As used herein, a "polymerase" refers to an enzyme that catalyzes the formation of polymers of nucleotides, that is, polynucleotides. A polymerase useful in the invention can be derived from any organism 5 or source, including animal, plant, bacterial and viral polymerases. A polymerase can be a DNA polymerase, RNA polymerase, or a reverse transcriptase capable of transcribing RNA into DNA.

10 As used herein the term "proofreading" describes the property of an enzyme where a nucleotide, such as, a mismatch nucleotide, can be removed by a 3'-to-5' exonuclease activity and replaced by, typically, a base-paired nucleotide.

15 As used herein, a "recombinant" polynucleotide refers to a polynucleotide that comprises sequence information from at least two different polynucleotides.

20 As used herein the term "related polynucleotides" means that regions or areas of the polynucleotides are identical and regions or areas of the polynucleotides are non-identical.

25 As used herein the term DNA "reassortment" is used herein to indicate a redistribution of sequence variations between substantially homologous but non-identical sequences.

As used herein the term "replicon" refers to a genetic unit of replication including a length of polynucleotide and its site for initiation of replication.

5

As used herein the term "sequence diversity" refers to the abundance of non-identical polynucleotides. The term "increasing sequence diversity in a population" means to increase the abundance of non-identical polynucleotides in a population.

10

As used herein the term "sequence variant" is used herein refers to a molecule (DNA, RNA polypeptide, and the like) with one or more sequence differences compared to a reference molecule. For example, the sum of the separate independent mismatch resolution events that occur throughout the heteroduplex molecule during the GRAMMR process results in reassortment of sequence information throughout that molecule. The sequence information will reassort in a variety of combinations to generate a complex library of "sequence variants".

15

As used herein the term "strand cleavage activity" or "cleavage" refers to the breaking of a phosphodiester bond in the backbone of the polynucleotide strand, as in forming a nick. Strand cleavage activity can be provided by an enzymatic agent, such agents include, but are not limited to CEL I, T4 endonuclease VII, T7 endonuclease I, S1 nuclease, BAL-31 nuclease, FEN1, cleavase, pancreatic DNase I, SP nuclease, mung bean nuclease, and nuclease P1; by a chemical agent, such

20

25

30

agents include, but are not limited to potassium permanganate, tetraethylammonium acetate, sterically bulky photoactivatable DNA intercalators, [Rh(bpy)₂(chrysi)]³⁺, osmium tetroxide with piperidine, 5 and hydroxylamine with piperidine; or by energy in the form of ionizing radiation, or kinetic radiation.

As used herein the term "sufficient time" refers to the period time necessary for a reaction or 10 process to render a desired product. For the present invention, the determination of sufficient time is well within the knowledge of those of ordinary skill in the art. It is noted that "sufficient time" can vary widely, depending on the desires of the practitioner, without 15 impacting on the functionality of the reaction, or the quality of the desired product.

As used herein the term "wild-type" means that a nucleic acid fragment does not contain any mutations. 20 A "wild-type" protein means that the protein will be active at a level of activity found in nature and typically will be the amino acid sequence found in nature. In an aspect, the term "wild type" or "parental sequence" can indicate a starting or reference sequence 25 prior to a manipulation of the invention.

In the polypeptide notation used herein, the left-hand direction is the amino terminal direction and the right-hand direction is the carboxy-terminal 30 direction, in accordance with standard usage and convention. Similarly, unless specified otherwise, the

left-hand end of single-stranded polynucleotide sequences is the 5' end; the left-hand direction of double-stranded polynucleotide sequences is referred to as the 5' direction. The direction of 5' to 3' addition of nascent 5 RNA transcripts is referred to as the transcription direction.

Detailed Description of the Invention

10 The present invention provides an in vitro method of making sequence variants from at least one heteroduplex polynucleotide wherein the heteroduplex has at least two non-complementary nucleotide base pairs, the method comprising: preparing at least one heteroduplex 15 polynucleotide; combining said heteroduplex polynucleotide with an effective amount of an agent or agents with exonuclease activity, polymerase activity and strand cleavage activity; and allowing sufficient time for the percentage of complementarity to increase, 20 wherein at least one or more variants are made.

Another aspect of the present invention is where the heteroduplex polynucleotides are circular, linear or a replicon.

25 Another aspect of the present invention is where the desired variants have different amounts of complementarity.

Another aspect of the present invention is where the exonuclease activity, polymerase activity, and strand cleavage activity is added sequentially, or concurrently.

5

Another aspect of the present invention provides the addition of ligase activity, provided by agents such as, T4 DNA ligase, E. coli DNA ligase, or Taq DNA ligase.

10

Another aspect of the present invention is where the strand cleavage activity is provided by an enzyme, such as, CEL I, T4 endonuclease VII, T7 endonuclease I, S1 nuclease, BAL-31 nuclease, FEN1, 15 cleavase, pancreatic DNase I, SP nuclease, mung bean nuclease, and nuclease P1; a chemical agent, such as, potassium permanganate, tetraethylammonium acetate, sterically bulky photoactivatable DNA intercalators, [Rh(bpy)₂(chrys)]³⁺, osmium tetroxide with piperidine, 20 and hydroxylamine with piperidine or a form of energy, such as, ionizing or kinetic radiation.

Another aspect of the present invention is where polymerase activity is provided by Pol beta.

25

Another aspect of the present invention is where both polymerase activity and 3' to 5' exonuclease activity is provided T4 DNA polymerase, T7 DNA polymerase, E. coli Pol 1, or Pfu DNA polymerase.

30

Another aspect of the present invention is where the agent with both polymerase activity and 5' to 3' exonuclease activity is E. coli Pol 1.

5 An embodiment of the present invention is where the effective amount of strand cleavage activity, and exonuclease activity/polymerase activity and ligase activity are provided by CEL I, T4 DNA polymerase, and T4 DNA ligase.

10 Another aspect of the present invention is where the effective amount of strand cleavage activity, and exonuclease activity/polymerase activity and ligase activity are provided by CEL I, T7 DNA polymerase, and T4 DNA ligase.

15 Another embodiment of the present invention provides an in vitro method of increasing diversity in a population of sequences, comprising, preparing at least 20 one heteroduplex polynucleotide; combining the heteroduplex polynucleotide with an effective amount of an agent or agents with 3' to 5' exonuclease activity, polymerase activity and strand cleavage activity; and allowing sufficient time for the percentage of 25 complementarity to increase, wherein diversity in the population is increased.

30 Another embodiment of the present invention provides a method of obtaining a polynucleotide encoding a desired functional property, comprising: preparing at least one heteroduplex polynucleotide; combining said

heteroduplex polynucleotide with an effective amount of an agent or agents with exonuclease activity, polymerase activity and strand cleavage activity; allowing sufficient time for the percentage of complementarity 5 between strands of the heteroduplex polynucleotide to increase, wherein diversity in the population is increased; and screening or selecting a population of variants for the desired functional property.

10 Another embodiment of the present invention provides a method of obtaining a polynucleotide encoding a desired functional property, comprising: preparing at least one heteroduplex polynucleotide; combining said heteroduplex polynucleotide with an effective amount of 15 an agent or agents with exonuclease activity, polymerase activity and strand cleavage activity; allowing sufficient time for the percentage of complementarity between strands of the heteroduplex polynucleotide to increase, wherein diversity in the population is increased; converting DNA to RNA; and screening or 20 selecting a population of ribonucleic acid variants for the desired functional property.

25 Yet another embodiment of the present invention provides a method of obtaining a polypeptide having a desired functional property, comprising: preparing at least one heteroduplex polynucleotide; combining said heteroduplex polynucleotide with an effective amount of an agent or agents with exonuclease activity, polymerase 30 activity and strand cleavage activity; allowing sufficient time for the percentage of complementarity

between strands of said heteroduplex polynucleotide to increase, converting said heteroduplex polynucleotide to RNA, and said RNA to a polypeptide; and screening or selecting a population of polypeptide variants for said 5 desired functional property.

Still another embodiment of the present invention provides a method of obtaining a polynucleotide encoding a desired functional property, comprising:

- 10 preparing at least one heteroduplex polynucleotide, where the heteroduplex is optionally, about 95%, 90%, 85%, 80%, or 75% identical, and about 1000KB, 10,000KB, or 100,000KB is size; combining said heteroduplex polynucleotide with an effective amount of an agent or 15 agents with exonuclease activity, polymerase activity and strand cleavage activity; allowing sufficient time for the percentage of complementarity between strands of the heteroduplex polynucleotide to increase, screening or selecting for a population of variants having a desired 20 functional property; denaturing said population of variants to obtain single strand polynucleotides; annealing said single strand polynucleotides to form at least one second heteroduplex polynucleotide; combining said second heteroduplex polynucleotide with an effective 25 amount of an agent or agents with exonuclease activity, polymerase activity and strand cleavage activity; and allowing sufficient time for the percentage of complementarity between strands of the heteroduplex polynucleotide to increase.

The present invention is directed to a method for generating an improved polynucleotide sequence or a population of improved polynucleotide sequences, typically in the form of amplified and/or cloned 5 polynucleotides, whereby the improved polynucleotide sequence(s) possess at least one desired phenotypic characteristic (e.g., encodes a polypeptide, promotes transcription of linked polynucleotides, binds a protein, improves the function of a viral vector, and the like) 10 which can be selected or screened for. Such desired polynucleotides can be used in a number of ways such as expression from a suitable plant, animal, fungal, yeast, or bacterial expression vector, integration to form a transgenic plant, animal or microorganism, expression of 15 a ribozyme, and the like.

GRAMMR provides for a process where heteroduplexed DNA strands are created by annealing followed by resolution of mismatches in an *in vitro* 20 reaction. This reaction begins with cleavage of one strand or the other at or near a mismatch followed by excision of mismatched bases from that strand and polymerization to fill in the resulting gap with nucleotides that are templated to the sequence of the 25 other strand. The resulting nick can be sealed by ligation to rejoin the backbone. The sum of the separate independent mismatch resolution events that occur throughout the heteroduplex molecule will result in reassortment of sequence information throughout that 30 molecule. The sequence information will reassort in a

variety of combinations to generate a complex library of sequence variants.

In one embodiment of GRAMMR, a library of mutants is generated by any method known in the art such as mutagenic PCR, chemical mutagenesis, etc. followed by screening or selection for mutants with a desired property. DNA is prepared from the chosen mutants. The DNAs of the mutants are mixed, denatured to single strands, and allowed to anneal. Partially complementary strands that hybridize will have non-base-paired nucleotides at the sites of the mismatches. Treatment with CEL I (Oleykowski *et al.*, 1998; Yang *et al.*, 2000), or a similar mismatch-directed activity, will cause nicking of one or the other polynucleotide strand 3' of each mismatch. (In addition, CEL I can nick 3' of an insertion/deletion resulting in reassortment of insertions/deletions.) The presence of a polymerase containing a 3'-to-5' exonuclease ("proofreading") activity (e.g., T4 DNA Pol) will allow excision of the mismatch, and subsequent 5'-to-3' polymerase activity will fill in the gap using the other strand as a template. A polymerase that lacks 5'-3' exonuclease activity and strand-displacement activity will fill in the gap and will cease to polymerize when it reaches the 5' end of DNA located at the original CEL I cleavage site, thus re-synthesizing only short patches of sequence. Alternatively, the length of the synthesized patches can be modulated by spiking the reaction with a polymerase that contains a 5'-3' exonuclease activity; this nick-translation activity can traverse a longer

region resulting in a longer patch of information transferred from the template strand. DNA ligase (e.g., T4 DNA ligase) can then seal the nick by restoring the phosphate backbone of the repaired strand. This process

5 can occur simultaneously at many sites and on either strand of a given heteroduplexed DNA molecule. The result is a randomization of sequence differences among input strands to give a population of sequence variants that is more diverse than the population of starting

10 sequences. These output polynucleotides can be cloned directly into a suitable vector, or they can be amplified by PCR before cloning. Alternatively, the reaction can be carried out on heteroduplexed regions within the context of a double-stranded circular plasmid molecule or

15 other suitable replicon that can be directly introduced into the appropriate host following the GRAMMR reaction. In another alternative, the output polynucleotides can be transcribed into RNA polynucleotides and used directly, for example, by inoculation of a plant viral vector onto

20 a plant, such as in the instance of a viral vector transcription plasmid. The resulting clones are subjected to a selection or a screen for improvements in a desired property. The overall process can then be repeated one or more times with the selected clones in an

25 attempt to obtain additional improvements.

If the output polynucleotides are cloned directly, there is the possibility of incompletely resolved molecules persisting that, upon replication in

30 the cloning host, could lead to two different plasmids in the same cell. These plasmids could potentially give

rise to mixed-plasmid colonies. If it is desired to avoid such a possibility, the output polynucleotide molecules can be grown in the host to allow replication/resolution, the polynucleotides isolated and 5 retransformed into new host cells.

In another embodiment, when sequence input from more than two parents per molecule is desired, the above procedure is performed in a cyclic manner before any 10 cloning of output polynucleotides. After GRAMMR treatment, the double stranded polynucleotides are denatured, allowed to anneal, and the mismatch resolution process is repeated. After a desired number of such cycles, the output polynucleotides can be cloned 15 directly, introduced into a suitable vector, or they can be amplified by PCR before cloning. The resulting clones are subjected to a selection or a screen for improvements in a desired property.

20 In another embodiment, a "molecular backcross" is performed to help eliminate the background of deleterious mutations from the desired mutations. A pool of desired mutants' DNA can be mixed with an appropriate ratio of wild-type DNA to perform the method. Clones can 25 be selected for improvement, pooled, and crossed back to wild-type again until there is no further significant change.

30 The efficiency of the process is improved by various methods of enriching the starting population for heteroduplex molecules, thus reducing the number of

unaltered parental-type output molecules. The mismatched hybrids can be affinity purified using aptamers, dyes, or other agents that bind to mismatched DNA. A preferred embodiment is the use of MutS protein affinity matrix

5 (Wagner *et al.*, Nucleic Acids Res. 23(19):3944-3948 (1995); Su *et al.*, Proc. Natl. Acad. Sci. (U.S.A.), 83:5057-5061(1986)) or mismatch-binding but non-cleaving mutants of phage T4 endonuclease VII (Golz and Kemper, Nucleic Acids Research, 1999; 27: e7).

10

In one embodiment, the procedure is modified so that the input polynucleotides consist of a single strand of each sequence variant. For example, single-stranded DNAs of opposite strandedness are produced from the 15 different parent sequences by asymmetric PCR to generate partially complementary single-stranded molecules. Annealing of the strands with one-another to make heteroduplex is performed as described in Example 1. Alternatively, single-stranded DNAs can be generated by 20 preferentially digesting one strand of each parental double-stranded DNA with Lambda exonuclease followed by annealing the remaining strands to one-another. In this embodiment, the annealing strands have no 100% complementary strand present with which to re-anneal. 25 Hence, there is a lower background of unmodified polynucleotides, that is, "parental polynucleotides" among the output polynucleotides leading to a higher efficiency of reassorting sequence variations. This increased efficiency will be particularly valuable in 30 situations where a screen rather than a selection is employed to test for the desired polynucleotides.

Another method for heteroduplex formation is to mix the double-stranded parent DNAs, denature to dissociate the strands, and allow the single-stranded 5 DNAs to anneal to one-another to generate a population of heteroduplexes and parental homoduplexes. The heteroduplexes can then be selectively enriched by a heteroduplex capture method such as those described above using MutS or a non-cleaving T4 endonuclease VII mutant. 10 Alternatively, the parental homoduplex molecules in the population may be cleaved by restriction enzymes that overlap with sites of mismatch such that they are not cleaved in the heteroduplex but are cleaved in the parental homoduplex molecules. Uncleaved heteroduplex 15 DNA can then be isolated by size fractionation in an agarose gel as was performed to generate full-length plasmid on full-length plasmid heteroduplex DNA molecules as describe in Example 6. Circularization of those full-length heteroduplexed plasmid molecules was then brought 20 about by incubation with DNA ligase.

In another embodiment, the parental, or input, double-stranded polynucleotides are modified by the addition of "clamp" sequences. One input polynucleotide 25 or pool of polynucleotides is amplified by PCR with the addition of a unique sequence in the 5' primer. The other input polynucleotide or pool is amplified by PCR with the addition of a unique sequence in the 3' primer. The clamp sequences can be designed to contain a unique 30 restriction enzyme site for the 5' end of the gene of interest and another for the 3' end such that, at the

step of cloning the products of the GRAMMR reassortment, only products with the 5' clamp from the first polynucleotide (or pool) and the 3' end from the second polynucleotide (or pool) will have appropriate ends for 5 cloning. Alternatively, the products of GRAMMR reassortment can be PCR amplified using the unique sequences of the 5' and 3' clamps to achieve a similar result. Hence, there is a lower background of unmodified polynucleotides, that is, "parental polynucleotides" 10 among the output polynucleotide clones leading to a higher efficiency of reassorting sequence variations. This increased efficiency will be particularly valuable in situations where a screen rather than a selection is employed to test for the desired polynucleotides. 15 Optionally, oligonucleotide primers can be added to the GRAMMR reaction that are complementary to the clamp primer sequences such that either parent can serve as the top strand, thus permitting both reciprocal heteroduplexes to participate in the mismatch-resolution 20 reaction.

Another method for generating cyclic heteroduplexed polynucleotides is performed where parental double-stranded DNAs have terminal clamp 25 sequences as described above where the single-stranded clamp sequences extending from one end of the heteroduplex are complementary to single-stranded clamp sequences extending from the other end of the heteroduplex. These complementary, single-stranded 30 clamps are allowed to anneal, thereby circularizing the heteroduplexed DNA molecule. Parental homoduplexes that

result from re-annealing of identical sequences have only one clamp sequence and therefore, no complementary single-stranded sequences at their termini with which circularization can occur. Additionally, a DNA 5 polymerase and a DNA ligase can be used to fill-in any gaps in the circular molecules and to seal the nicks in the backbone, respectively, to result in the formation of a population of covalently-closed circular heteroduplex molecules. As the covalently-closed circular 10 heteroduplex molecules will not dissociate into their component strands if subjected to further denaturating conditions, the process of denaturation, circularization, and ligation can be repeated to convert more of the linear double-stranded parental duplexes into closed into 15 closed circular heteroduplexes.

In another embodiment, a region of a single-stranded circular phagemid DNA can be hybridized to a related, but non-identical linear DNA, which can then be 20 extended with a polymerase such as T7 DNA polymerase or T4 DNA polymerase plus T4 gene 32 protein, then ligated at the resulting nick to obtain a circular, double-stranded molecule with heteroduplexed regions at the sites of differences between the DNAs. GRAMMR can then 25 be carried out on this molecule to obtain a library of sequence-reassorted molecules.

Alternately, two single-stranded circular phagemid DNAs of opposite strand polarity relative to the 30 plasmid backbone, and parent gene sequences that are the target of the reassortment are annealed to one and other.

A region of extensive mismatch will occur where the phage f1 origin sequences reside. Upon GRAMMR treatment, however, this region of extensive mismatch can revert to either parental type sequence restoring a functional f1 origin. These double stranded molecules will also contain mismatch regions at the sites of differences between the strands encoding the parent genes of interest. GRAMMR can then be carried out on this molecule to obtain a library of sequence re-assorted molecule.

As discussed in the preceding paragraphs, the starting DNA or input DNA can be of any number of forms. For example, input DNA can be full-length, single stranded and of opposite sense, as is taught in Example 1. Alternatively, the input DNA can also be a fragment of the full-length strand. The input DNAs can be double-stranded, either one or both, or modified, such as by, methylation, phosphorothiolate linkages, peptide-nucleic acid, substitution of RNA in one or both strands, or the like. Either strand of a duplex can be continuous along both strands, discontinuous but contiguous, discontinuous-with overlaps, or discontinuous with gaps.

GRAMMR can also be applied to DNA fragmentation and reassembly-based DNA shuffling schemes. For instance, in methods where gene fragments are taken through cycles of denaturation, annealing, and extension in the course of gene reassembly, GRAMMR can be employed as an intermediate step.

In one such embodiment, the DNA from a gene, or pool of mutants' genes is fragmented by enzymatic, mechanical or chemical means, and optionally a size range of said fragments is isolated by a means such as

5 separation on an agarose gel. The starting polynucleotide, such as a wild-type, or a desired variant, or a pool thereof, is added to the fragments and the mixture is denatured and then allowed to anneal. The annealed polynucleotides are treated with a polymerase to

10 fill in the single stranded gaps using the intact strand as a template. The resulting partially complementary double strands will have non-base-paired nucleotides at the sites of the mismatches. Treatment with CEL I (Oleykowski *et al.*, 1998; Yang *et al.*, 2000) will cause

15 nicking of one or the other polynucleotide strand 3' of each mismatch. Addition of a polymerase containing a 3'-to-5' exonuclease that provides proofreading activity, such as, DNA Pol I, T4 DNA Pol I, will allow excision of the mismatch, and subsequent 5'-to-3' polymerase activity

20 will fill in the gap using the other strand as a template. A DNA ligase, such as, T4 DNA Ligase, can then seal the nick by restoring the phosphate backbone of the repaired strand. The result is a randomization of sequence variation among input strands to give output

25 strands with potentially improved properties. These output polynucleotides can be cloned directly into a suitable vector, or they can be amplified by PCR before cloning. The resulting clones are subjected to a selection or a screen for improvements in a desired

30 property.

In one such embodiment, the DNA from a pool of mutants' genes is fragmented by enzymatic, mechanical or chemical means, or fragments are generated by limited extension of random oligonucleotides annealed to parental

5 templates (United States Patent 5,965,408), and optionally a size range of said fragments is isolated by a means such as separation on an agarose gel. The mixture is denatured and then allowed to anneal. The annealed polynucleotides are optionally treated with a

10 polymerase to fill in the single stranded gaps. The resulting partially complementary double-strand fragments will have non-base paired nucleotides at the sites of the mismatches. Treatment with CEL I (Oleykowski *et al.*, 1998; Yang *et al.*, 2000) will cause nicking of one or the

15 other polynucleotide strand 3' of each mismatch. The activity of a polymerase containing a 3'-to-5' exonuclease ("proofreading") activity, such as T4 DNA Polymerase, will allow excision of the mismatch, and subsequent 5'-to-3' polymerase activity will fill in the

20 gap using the other strand as a template. Optionally, DNA ligase, such as, T4 DNA Ligase, can then seal the nick by restoring the phosphate backbone of the repaired strand. The result is a randomization of sequence variation among input strands to give output strands with

25 potentially improved properties. Subsequent rounds of denaturing, annealing, and GRAMMR treatment allows gene reassembly. PCR can be used to amplify the desired portion of the reassembled gene. These PCR output polynucleotides can be cloned into a suitable vector.

30 The resulting clones are subjected to a selection or a screen for the desired functional property.

Another embodiment of the present invention provides starting with a continuous scaffold strand to which fragments of another gene or genes anneal. The 5 flaps and gaps are trimmed and filled as is described in Coco, et al., *Nature Biotech* 19 (01) 354; United States Patent 6,319,713, and GRAMMR is performed. In this process, GRAMMR would bring about further sequence reassortment by permitting transfer of sequence 10 information between the template strand and the strand resulting from flap and gap trimming and ligation. This method provides the benefits of incorporating specific sequence patches into one continuous strand followed by GRAMMR of residues that mismatch with the scaffold. By 15 annealing many fragments simultaneously to the same sequence or gene, many individual sites can be addressed simultaneously, thereby allowing reassortment of multiple sequences or genes at once. Unlike the method disclosed by Coco, et al., in the present embodiment, the scaffold 20 is not degraded, rather the duplex can be directly cloned, or amplified by PCR prior to cloning. Exhaustive mismatch resolution will result in a perfectly duplexed DNA. Partial mismatch resolution will result in essentially two different reassorted products per duplex.

25

As can be appreciated from the present disclosure, GRAMMR can also be applied to a variety of methods that include the annealing of related DNAs as a step in their process. For example, many site-directed 30 mutagenesis protocols call for the annealing of mutant-encoding DNA molecules to a circular DNA in single-

stranded form, either phagemid or denatured plasmid. These DNAs are then extended with a polymerase, followed by treatment with ligase to seal the nick, with further manipulation to remove the parental sequence, leaving the 5 desired mutation or mutations incorporated into the parental genetic background. Though these protocols are generally used to incorporate specific mutations into a particular DNA sequence, it is feasible that the GRAMMR process can be applied to the heteroduplexed molecules 10 generated in such a process to reassort sequence variations between the two strands, thereby resulting in a diverse set of progeny with reassorted genetic variation.

15 Another embodiment provides for a sequential round of reassortment on a particular region. For example, DNA fragments are annealed to a circular single-strand phagemid DNA, and GRAMMR is performed. The fragments can be treated in order to prevent them from 20 being physically incorporated into the output material. For example, they can be terminated at the 3' end with di-deoxy residues making them non-extendible. Multiple rounds of reassortment can be performed, but only modified molecules from the original input single 25 stranded DNA clone will be recovered. The consequence will be that the DNA fragments used in this reassortment will contribute only sequence information to the final product and will not be physically integrated into the final recoverable product.

In instances where it is desired to resolve only sites of significant mismatch, that is patches of more than about 1 to 3 mismatches, S1 nuclease can be used. S1 nuclease is an endonuclease specific for 5 single-stranded nucleic acids. It can recognize and cleave limited regions of mismatched base pairs in DNA:DNA or DNA:RNA duplexes. A mismatch of at least about 4 consecutive base pairs is generally required for recognition and cleavage by S1 nuclease. Mismatch 10 resolution will not occur if both strands are cleaved, so the DNA must be repaired after the first nick and before the counter-nick. Other nucleases may be preferable for specifically tuning cleavage specificity according to sequence, sequence context, or size of mismatch.

15

In addition, other means of addressing mismatched residues, such as chemical cleavage of mismatches may be used. Alternatively, one can choose to subject the strands of heteroduplexed DNA to random 20 nicking with an activity such as that exhibited by DNaseI or an agent that cleaves only in duplexed regions. If nick formation occurs in a region of identity between the two genes, the DNA ligase present in the reaction will seal the nick with no net transfer of sequence 25 information. However, if nick formation occurs near a site of mismatch, the mismatched bases can be removed by 3'-5' exonuclease and the gap filled in by polymerase followed by nick sealing by ligase. Alternatively, application of nick-translation through regions of 30 heterogeneity can bring about sequence reassortment. These processes, though not directed exclusively by the

RECORDED BY COMPUTER

mismatch status of the DNA, will serve to transfer sequence information to the repaired strand, and thus result in a reassorted sequence.

5 GRAMMR can be used for protein, peptide, or aptamer display methods to obtain recombination between library members that have been selected. As fragmentation of the input DNAs is not required for GRAMMR, it may be possible to reassort sequence
10 information between very small stretches of sequence. For instance, DNAs encoding small peptides or RNA aptamers that have been selected for a particular property such as target binding can be reassorted. For annealing to occur between the selected DNA molecules,
15 some level of sequence homology should be shared between the molecules, such as at the 5' and 3' regions of the coding sequence, in regions of the randomized sequence segment that bear similarity because of similar binding activities, or through the biasing of codon wobble-base
20 identity to a particular set of defaults.

Manipulation of the reaction temperature at which GRAMMR is conducted can be useful. For example, lower temperatures will help to stabilize heteroduplexes
25 allowing GRAMMR to be performed on more highly mismatched substrates. Likewise, additives that affect base-pairing between strands, such as salts, PEG, formamide, etc, can be used to alter the stability of the heteroduplex in the GRAMMR, thereby affecting the outcome of the reaction.

In another embodiment, the mismatched double stranded polynucleotides are generated, treated with a DNA glycosylase to form an apurinic or apyrimidinic site, (that is an "AP site") an AP endonuclease activity to 5 cleave the phosphodiester bond, deoxyribulose phosphodiesterase to remove the deoxyribose-phosphate molecules, DNA polymerase β or other DNA polymerase to add a single nucleotide to the 3' end of the DNA strand at the gap, and DNA ligase to seal the gap. The result 10 is a reassortment of sequence variations between input strands to give output strands with potentially improved properties. These output polynucleotides can be cloned directly into a suitable vector, or they can be amplified by PCR before cloning. The resulting clones are 15 subjected to a selection or a screen for improvements in a desired property.

Another embodiment provides for zonal mutagenesis by GRAMMR, that is, random or semi-random 20 mutations at, and in the immediate vicinity of, mismatched residues using nucleotide analogues that have multiple base-pairing potential. This provides for concentration of essentially random mutagenesis at a particular point of interest, and adds another benefit to 25 the present invention. Similar genes with slightly different functions, for example, plant R-genes, enzymes, or the like, will exhibit moderate sequence differences between them in regions that will be important for their own particular activities. Genes that express these 30 activities, such as different substrates, binding partners, regulatory sites, or the like, should have

heterogeneity in the regions that govern these functions. Since it is known that the specificity of such functions is associated with these amino acids and their neighbors, GRAMMR mutagenesis might serve to both reassort sequence variation among genes and also direct random mutagenesis to these regions to drive them further and faster evolutionarily, while not disturbing other sequences, such as structural framework, invariant residues, and other such important sites, that are potentially less tolerant to randomization.

Different enzymes with distinct functions will not differ just in the operative regions, such as active sites, regulatory sites, and the like. They are likely to have other differences from one another that arise through genetic drift. Further randomization in the locales of such changes might therefore be considered neutral, minimally important, or deleterious to the outcome of a mutagenesis experiment. In order to direct the random mutagenesis away from such inconsequential sites, and toward sites that might present a better result for random mutagenesis, such as the active site of an enzyme, the codon usage bias of the genes could be manipulated to decrease or increase the overall level of nucleotide complementarity in those regions. If regions of greater complementarity are less susceptible to GRAMMR than regions of lesser complementarity, then the degree of GRAMMER-directed zonal random mutagenesis at a given site can be modulated.

In another embodiment, after heteroduplex molecules are formed, an enzyme with a 3' to 5' exonuclease activity is added such that one strand of each end of the heteroduplex is digested back. At a 5 point at which, on average, a desired amount of 3' to 5' digestion has occurred, dNTPs are added to allow the 5' to 3' polymerase activity from the same or an additional enzyme to restore the duplex using the opposite strand as a template. Thus mismatches in the digested regions are 10 resolved to complementarity. Optionally, the resultant duplexes are purified, denatured and then allowed to anneal. The process of digestion, then polymerization is repeated resulting in new chimeric sequences. Additional cycles of the process can be performed as desired.

15 Output duplex molecules are cloned and tested for the desired functional property. This process requires no fragmentation and reassembly. In addition, this process requires no endonucleolytic cleavages.

20 In another embodiment, after the heteroduplex molecules are formed, an enzyme with a 5' to 3' exonuclease activity, such as, T7 Gene6 Exonuclease as disclosed in Enger, MJ and Richardson, CC, J Biol Chem 258(83)11197, is added such that one strand of each end 25 of the heteroduplex is digested. At a point at which, on average, a desired amount of 5' to 3' digestion has occurred, the reaction is stopped and the exonuclease inactivated. Oligonucleotide primers complementary to the 5' and 3' ends of the target polynucleotides are 30 added and annealed. A DNA polymerase, such as, T4 DNA Polymerase, a DNA ligase and dNTPs are added to allow the

5' to 3' polymerase activity to extend the primers and restore the duplex using the opposite strand as a template, with ligase sealing the nick. Thus mismatches in the digested regions are resolved to complementarity.

5 Optionally, the resultant duplexes are purified, denatured and then allowed to anneal. The process of digestion then polymerization is repeated resulting in new chimeric sequences. Additional cycles of the process can be performed as desired. Output duplex molecules are
10 cloned and tested for the desired functional property. This process requires no fragmentation and reassembly. In addition, this process requires no endonucleolytic cleavages.

15 In the current invention the random reassortment occurs in an *in vitro* DNA mismatch-resolution reaction. This method does not require any steps of "gene reassembly" that serve as the foundation for the earlier mutation reassortment ("shuffling")
20 methods. Instead, it is based upon the ability of a reconstituted or artificial DNA mismatch resolving system to transmit sequence variations from one or more strands of DNA into another DNA strand by hybridization and mismatch resolution *in vitro*.

25 In general, standard techniques of recombinant DNA technology are described in various publications, e.g., (Ausubel, 1987; Ausubel, 1999; Sambrook *et al.*, 1989), each of which is incorporated herein in their
30 entirety by reference. Polynucleotide modifying enzymes were used according to the manufacturers recommendations.

If desired, PCR amplimers for amplifying a predetermined DNA sequence may be chosen at the discretion of the practitioner.

5 It is noted that each of the activities taught in the present invention that are involved in the GRAMMR reaction can be interchanged with a functional equivalent agent with similar activity, and that such changes are within the scope of the present invention. For instance,
10 as was indicated in Example 2, Taq DNA ligase could substitute for T4 DNA ligase. Other ligases can be substituted as well, such as *E. coli* DNA ligase. Likewise, as shown in Examples 2 and 8, respectively, Pfu polymerase and T7 DNA polymerase can be substituted for
15 T4 DNA polymerase. Other enzymes with appropriate exonuclease activity with or without associated polymerase can function in place of any of these enzymes for the exonuclease activity needed for the GRAMMR reaction. In a similar way, any polymerase with
20 functionally equivalent activity to those demonstrated to work for GRAMMR can be used for substitution. These include *E. coli* Pol 1, the Klenow fragment of *E. coli* Pol 1, polymerase beta, among many others.

25 Strand cleavage may be brought about in a number of ways. In addition to CEL I, a number of functionally equivalent, and potentially homologous activities found in extracts from a variety of plant species (Oleykowsky, Nucleic Acids Res 1998;26:4597-602)
30 may be used. Other mismatch-directed endonucleases such as T4 endonuclease VII, T7 endonuclease I, and SP

nuclease (Oleykowsky, Biochemistry 1999; 38: 2200-5) may be used. Other nucleases which attack single stranded DNA can be used, such as S1 nuclease, FEN1, cleavase, mung bean nuclease, and nuclease P1. Enzymes that make 5 random cleavage events in DNA, such as pancreatic DNase I may also be substituted for the strand cleaving activity in GRAMMR. A number of methods for bringing about strand cleavage through other means are also envisioned. These include potassium permanganate used with 10 tetraethylammonium acetate, the use of sterically bulky photoactivatable DNA intercalators such as [Rh(bpy)2(chrysi)]3+, osmium tetroxide with piperidine alkaloid, and hydroxylamine with piperidine alkaloid, as well as the use of radiation energy to bring about strand 15 breakage.

CEL I is a mismatch endonuclease isolated from celery. The use of CEL I in a diagnostic method for the detection of mutations in targeted polynucleotide 20 sequences, in particular, those associated with cancer, is disclosed in U.S. Patent No. 5,869,245. Methods of isolating and preparing CEL I are also disclosed in this patent. However, there is no disclosure in this patent relating to the use of CEL I in DNA sequence 25 reassortment.

The nucleic acid molecules that encode CEL I are disclosed in PCT Application Publication No. WO 01/62974 A1. As with U.S. Patent No. 5,869,245, the use 30 of CEL I in a diagnostic method for the detection of mutations in targeted polynucleotide sequences associated

with cancer is disclosed. Also similarly, there is no disclosure relating to the use of CEL I in DNA reassortment.

5 The reactivity of Endonuclease VII of phage T4 with DNA-loops of eight, four, or one nucleotide, or any of 8 possible base mismatches *in vitro* is disclosed in "Endonuclease VII of Phage T4 Triggers Mismatch Correction *in Vitro*" Solaro, et al., *J Mol Biol* 10 230(93)868. The publication reports a mechanism where Endonuclease VII introduces double stranded breaks by creating nicks and counternicks within six nucleotides 3' of the mispairing. The publication discloses that a time delay between the occurrence of the first nick and the 15 counternick was sufficient to allow the 3'-5' exonuclease activity of gp43 to remove the mispairing and its polymerase activity to fill in the gap before the occurrence of the counternick. Nucleotides are erased from the first nick, which is located 3' of the mismatch 20 on either strand and stops 5' of the mismatch at the first stable base-pair. The polymerase activity proceeds in the 5' to 3' direction towards the initial nick, which is sealed by DNA ligase. As a result, very short repair tracks of 3 to 4 nucleotides extend across the site of 25 the former mismatch. The publication concludes with a discussion regarding the various activities Endonuclease VII may have within phage T4. However, the publication does not disclose any practical utility for Endonuclease VII outside of phage T4, and there is no disclosure 30 regarding its applicability in DNA reassortment.

A method for creating libraries of chimeric DNA sequences *in vivo* in *Escherichia coli* is disclosed in *Nucleic Acids Research*, 1999, Vol 27, No. 18, e18, Volkov, A. A., Shao, Z., and Arnold, F.H. The method 5 uses a heteroduplex formed *in vitro* to transform *E. coli* where repair of regions of non-identity in the heteroduplex creates a library of new, recombined sequences composed of elements of each parent. Although the publication discloses the use of this method as a 10 convenient addition to existing DNA recombination methods, that is, DNA shuffling, the disclosed method is limited to the *in vivo* environment of *E. coli*. The publication states that there is more than one mechanism available for mismatch repair in *E. coli*, and that the 15 'long patch' repair mechanism, which utilizes the MutS/L/H enzyme system, was probably responsible for the heteroduplex repair.

Cited References

20

1. Arkin, A.P. and Youvan, D.C. (1992) An algorithm for protein engineering: simulations of recursive ensemble mutagenesis. *Proc Natl Acad Sci U S A*, **89**, 7811-7815.
- 25 2. Ausubel, F.M. (1987) *Current protocols in molecular biology*. Published by Greene Pub. Associates and Wiley-Interscience : J. Wiley, New York.
3. Ausubel, F.M. (1999) *Short protocols in molecular biology : a compendium of methods from Current protocols in molecular biology*. Wiley, New York.

30

4. Barnes, W.M. (1994) PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proc Natl Acad Sci U S A*, **91**, 2216-2220.

5

5. Bartel, D.P. and Szostak, J.W. (1993) Isolation of new ribozymes from a large pool of random sequences. *Science*, **261**, 1411-1418.

10 6. Cadwell, R.C. and Joyce, G.F. (1992) Randomization of genes by PCR mutagenesis. *PCR Methods Appl*, **2**, 28-33.

7. Calogero, S., Bianchi, M.E. and Galizzi, A. (1992) In vivo recombination and the production of hybrid genes.

15 *FEMS Microbiol Lett*, **76**, 41-44.

8. Caren, R., Morkeberg, R. and Khosla, C. (1994) Efficient sampling of protein sequence space for multiple mutants. *Biotechnology (N Y)*, **12**, 517-520.

20

9. Delagrange, S., Goldman, E.R. and Youvan, D.C. (1993) Recursive ensemble mutagenesis. *Protein Eng*, **6**, 327-331.

25 10. Delagrange, S. and Youvan, D.C. (1993) Searching sequence space to engineer proteins: exponential ensemble mutagenesis. *Biotechnology (N Y)*, **11**, 1548-1552.

11. Goldman, E.R. and Youvan, D.C. (1992) An algorithmically optimized combinatorial library screened by digital imaging spectroscopy. *Biotechnology (N Y)*, **10**,

30 1557-1561.

12. Gram, H., Marconi, L.A., Barbas, C.F.d., Collet, T.A., Lerner, R.A. and Kang, A.S. (1992) In vitro selection and affinity maturation of antibodies from a naive combinatorial immunoglobulin library. *Proc Natl Acad Sci U S A*, **89**, 3576-3580.

13. Hayashi, N., Welschof, M., Zewe, M., Braunagel, M., Dubel, S., Breitling, F. and Little, M. (1994) Simultaneous mutagenesis of antibody CDR regions by overlap extension and PCR. *Biotechniques*, **17**, 310, 312, 314-315.

14. Hermes, J.D., Blacklow, S.C. and Knowles, J.R. (1990) Searching sequence space by definably random mutagenesis: improving the catalytic potency of an enzyme. *Proc Natl Acad Sci U S A*, **87**, 696-700.

15. Holland, J.H. (1992) *Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence*. MIT Press, Cambridge, Mass.

16. Ji, G. and Silver, S. (1992) Regulation and expression of the arsenic resistance operon from *Staphylococcus aureus* plasmid pI258. *J Bacteriol*, **174**, 3684-3694.

17. Kauffman, S.A. (1993) *The origins of order : self-organization and selection in evolution*. Oxford University Press, New York.

18. Marton, A., Delbecchi, L. and Bourgaux, P. (1991) DNA nicking favors PCR recombination. *Nucleic Acids Res*, **19**, 2423-2426.

5 19. Meyerhans, A., Vartanian, J.P. and Wain-Hobson, S. (1990) DNA recombination during PCR. *Nucleic Acids Res*, **18**, 1687-1691.

10 20. Nissim, A., Hoogenboom, H.R., Tomlinson, I.M., Flynn, G., Midgley, C., Lane, D. and Winter, G. (1994) Antibody fragments from a 'single pot' phage display library as immunochemical reagents. *Embo J*, **13**, 692-698.

15 21. Oleykowski, C.A., Bronson Mullins, C.R., Godwin, A.K. and Yeung, A.T. (1998) Mutation detection using a novel plant endonuclease. *Nucleic Acids Res*, **26**, 4597-4602.

20 22. Oliphant, A.R., Nussbaum, A.L. and Struhl, K. (1986) Cloning of random-sequence oligodeoxynucleotides. *Gene*, **44**, 177-183.

25 23. Sambrook, J., Maniatis, T. and Fritsch, E.F. (1989) *Molecular cloning : a laboratory manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

30 24. Stemmer, W.P. (1994a) DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc Natl Acad Sci U S A*, **91**, 10747-10751.

25. Stemmer, W.P. (1994b) Rapid evolution of a protein in vitro by DNA shuffling. *Nature*, **370**, 389-391.

26. Stemmer, W.P., Morris, S.K. and Wilson, B.S. (1993) 5 Selection of an active single chain Fv antibody from a protein linker library prepared by enzymatic inverse PCR. *Biotechniques*, **14**, 256-265.

27. Winter, G., Griffiths, A.D., Hawkins, R.E. and 10 Hoogenboom, H.R. (1994) Making antibodies by phage display technology. *Annu Rev Immunol*, **12**, 433-455.

28. Yang, B., Wen, X., Kodali, N.S., Oleykowski, C.A., Miller, C.G., Kulinski, J., Besack, D., Yeung, J.A., 15 Kowalski, D. and Yeung, A.T. (2000) Purification, cloning, and characterization of the CEL I nuclease. *Biochemistry*, **39**, 3533-3541.

20 The following non-limiting examples are provided to illustrate the present invention.

EXAMPLE 1**Cleavage of Mismatched DNA Substrate by CEL I**

5

This example teaches the preparation of CEL I enzyme and its use in the cleavage of mismatched DNA substrate.

10 CEL I enzyme was prepared from celery stalks using the homogenization, ammonium sulfate, and Concanavalin A-Sepharose protocol described by Yang et al. (Biochemistry, 39:3533-3541 (2000), incorporated herein by reference. A 1.5 kg sample of chilled celery
15 stalks was homogenized with a juice extractor. One liter of juice was collected, adjusted to 100 mM Tris-HCL, pH 7.7 with 100 micromolar phenylmethylsulfonyl fluoride (PMSF), and filtered through two layers of miracloth. Solid $(\text{NH}_4)_2\text{SO}_4$ was slowly added to 25% saturation while
20 stirring on ice. After 30 minutes, the suspension was centrifuged at 27,000g for 1.5 hours at 4°C. The supernatants were collected and adjusted with solid $(\text{NH}_4)_2\text{SO}_4$ to 80% saturation while stirring on ice followed by centrifugation at 27,000g for 2 hours. The pellets
25 were re-suspended in buffer B (0.1 M Tris-HCL, pH 7.7, 0.5 M KCl, 100 micromolar PMSF) and dialyzed against the same buffer.

Conconavalin A (ConA) Sepharose affinity
30 chromatography was performed by first incubating the dialyzed sample with 2 ml of ConA resin overnight with

gentle agitation. The ConA resin was then packed into a 0.5 cm diameter column and washed with several column volumes of buffer B. Elution was performed using 0.3 M alpha-methyl-mannoside in buffer B. Fractions were 5 collected in 1 ml aliquots. Fractions were assayed for mismatch cleavage activity on a radiolabeled mismatch substrate by incubating 0.1 microliter of each fraction with the mismatched probe in buffer D (20 mM Tris-HCL, pH 7.4, 25 mM KCL, 10 mM MgCl₂) for 30 minutes at 45°C as 10 described by Oleykowski et al. (Nucleic Acids Research 26: 4597-4602 (1998), incorporated herein by reference. Reaction products were visualized by separation on 10% TBE-PAGE gels containing 7% urea (Invitrogen), followed by autoradiography. Aliquots of the CEL I fractions 15 having mismatch cleavage activity were stored frozen at -20°C. A series of five-fold dilutions of CEL I fraction #5 were then analyzed for mismatch cleavage of radiolabeled mismatch substrate. Reactions were performed either in buffer D, New England BioLabs (NEB) 20 T4 DNA ligase buffer (50 mM Tris-HCL, pH 7.5, 10 mM MgCl₂, 10 mM dithiothreitol (DTT), 1 mM ATP, 25 microgram/ml BSA), or Gibco/BRL T4 DNA ligase buffer (50 mM Tris-HCL, pH 7.6, 10mM MgCl₂, 1 mM DTT, 1 mM ATP, 5% (w/v) polyethylene glycol-8000). Reaction products 25 were visualized as above. Cleavage activity in buffer D and in NEB T4 DNA ligase buffer were found to be roughly equivalent, whereas cleavage in the PEG-containing Gibco/BRL ligase buffer was enhanced by five to ten-fold compared to the other buffers.

Additional analysis of CEL I activity was carried out using defined heteroduplex DNAs from two different Green Fluorescent Protein (GFP) genes as substrate. This GFP heteroduplex substrate was prepared 5 by annealing single stranded DNAs corresponding to cycle 3 GFP on the sense strand and wild-type GFP on the antisense strand. The single-stranded DNAs had been synthesized by asymmetric PCR and isolated by agarose gel electrophoresis. After annealing by heating to 90°C and 10 cooling in the presence of 1X NEB restriction enzyme buffer 2 (10 mM Tris-HCL, pH 7.9, 10 mM MgCl₂, 50 mM NaCl, 1 mM dithiothreitol), the heteroduplex DNA was isolated by agarose gel electrophoresis followed by excision of the heteroduplex band and extraction using 15 Qiaquick DNA spin columns. A total of twenty eight mismatches, one or two nucleotides in length, occur throughout the length of the heteroduplex molecule. The distribution of the mismatches ranges from small clusters of several mismatches separated by one or two nucleotides 20 to mismatches separated by more than thirty base pairs on either side.

A series of three-fold dilutions of CEL I in 1X NEB T4 DNA ligase buffer were prepared and one microliter 25 aliquots of each were incubated in two separate series of 10 microliter reactions, each containing as substrate either 0.5 microgram of a supercoiled plasmid preparation or one hundred nanograms of the cycle3/wild-type GFP heteroduplex. All reactions took place in 1X NEB T4 DNA 30 ligase buffer. Reactions were incubated at 45°C for 30

minutes and run on 1.5% TBE-agarose gel in the presence of ethidium bromide.

Treatment of the supercoiled plasmid

5 preparation with increasing amounts of CEL I resulted in the conversion of supercoiled DNA to nicked circular, then linear molecules, and then to smaller fragments of DNA of random size. Treatment of the mismatched GFP substrate with the CEL I preparation resulted in the 10 digestion of the full-length heteroduplex into laddered DNA bands which are likely to represent cleavage on opposite DNA strands in the vicinity of clusters of mismatches. Further digestion resulted in the conversion of the mismatched GFP substrate to smaller DNAs that may 15 represent a limit digest of the heteroduplex DNA by the CEL I preparation.

EXAMPLE 2

20 **Conservation of Full Length GFP Gene
with Mismatch Resolution Cocktails**

25 This example teaches various mismatch resolution cocktails that conserve the full length GFP Gene.

Mismatched GFP substrate was treated with various concentrations of CEL I in the presence of cocktails of enzymes that together constitute a synthetic 30 mismatch resolution system. The enzymes used were CEL I, T4 DNA polymerase, Taq DNA polymerase and T4 DNA ligase.

CEL I activity should nick the heteroduplex 3' of mismatched bases. T4 DNA polymerase contains 3'-5' exonuclease for excision of the mismatched base from the nicked heteroduplex. T4 DNA polymerase and Taq DNA 5 polymerase contain DNA polymerase capable of filling the gap. T4 DNA ligase seals the nick in the repaired molecule. Taq DNA polymerase also has 5' flap-ase activity.

10 Matrix experiments were performed to identify the reaction conditions that would serve to resolve mismatches in the GFP heteroduplex substrate. In one experiment, cycle 3/wild-type GFP heteroduplex was incubated in a matrix format with serial dilutions of CEL 15 I fraction number five (described above) at eight different concentrations. Each reaction contained 100 nanograms of heteroduplex substrate and 0.2 microliters of T4 DNA ligase (Gibco BRL) in 1X NEBT4 DNA ligase buffer and dNTPs at 250 micromolar each, in a reaction 20 volume of 10 microliters. In all, the matrix contained 96 individual reactions. One full set of reactions was incubated at room temperature for 30 minutes while another full set was incubated at 37°C for 30 minutes.

25 After incubation, PCR was used to amplify the GFP gene from each reaction. Aliquots from each PCR were then digested with *Hind*III and *Hpa*I and electrophoresed on 3% agarose gels with ethidium bromide. Only cycle 3 GFP has a *Hind*III site and only wild-type encodes a *Hpa*I 30 site.

If DNA mismatch resolution occurred at either the *Hind*III or *Hpa*I mismatched sites, then a proportion of the PCR product would be expected to contain both sites, yielding a novel band. The band was observed in 5 all samples, including the negative control samples that had neither CEL I, nor T4 DNA polymerase, nor Taq DNA polymerase. The results suggested that a basal level of background recombination may have occurred at some point in the experiment other than in the GRAMMR reaction; 10 possibly in the PCR step. PCR-mediated recombination is known to occur at some frequency between related sequences during amplification [reference Paabo, et al., DNA damage promotes jumping between templates during enzymatic amplification. *J Biol Chem* 265(90)4718-4721].

15

In another experiment, 200 nanograms of cycle 3/wild-type GFP heteroduplex was treated with CEL I and T4 DNA polymerase in various concentrations along with 2.5 units of Taq DNA polymerase in the presence or 20 absence of T4 DNA ligase (0.2 units; Gibco BRL). Each reaction contained 1X NEB T4 DNA ligase buffer with 0.05 mM each dNTP in a final volume of 20 microliters. Reactions were incubated for 30 minutes at 37°C and 10 microliters were run on a 2% TBE-agarose gel in the 25 presence of ethidium bromide. Results showed that in the presence of DNA ligase, but in the absence of T4 DNA polymerase, increasing amounts of CEL I caused greater degradation of the heteroduplexed DNA, but that this effect could be counteracted by increasing the amount of 30 T4 DNA polymerase in the reaction. These results indicated that the various components of the complete

reaction could act together to conserve the integrity of the full-length gene through DNA mismatch resolution.

Another matrix experiment was conducted to 5 expand on these results and to identify additional conditions for DNA mismatch resolution for this synthetic system. 60 nanograms of cycle3/wild-type GFP heteroduplex were treated with CEL I and T4 DNA polymerase at various concentrations in the presence of 10 2.5 units of Taq DNA polymerase and 0.2 units of T4 DNA ligase in 1 X NEB T4 DNA ligase buffer containing 0.5 mM of each dNTP in a reaction volume of 10 microliters. Each set of reactions was incubated for 1 hour at either 20°C, 30°C, 37°C, or at 45°C. All reactions were then run 15 on a 1.5% TBE-agarose gels in the presence of ethidium bromide. The results showed that the GFP heteroduplex was cleaved into discrete fragments by the CEL I preparation alone. The success of DNA mismatch resolution was initially gauged by the degree to which 20 the apparent full-length integrity of the GFP sequence was maintained by the other components of the mismatch resolution system in the presence of CEL I. Conditions of enzyme concentration and temperature were identified that conserved a high proportion of the DNA as full-length molecules in this assay. Namely, one microliter 25 of the CEL I fraction five preparation (described in Example 1) with one microliter (1 unit) of the T4 DNA polymerase in the presence of the other reaction components which were held constant in the experiment. 30 It was found that as the reaction temperature increased, the degradative activity of CEL I increased accordingly.

Furthermore, it was shown that the other components of the repair reaction acted to conserve the integrity of the full-length DNA at 20°C, 30°C, and 37°C, but was remarkably less efficient at conserving the full-length 5 DNA at 45°C. From these results, we concluded that under these experimental conditions, incubation at 45°C was not optimal for the process of GRAMMR, and that incubation at 20°C, 30°C, and 37°C were permissible.

10 Another experiment was performed in which alternative enzymes were used for the DNA mismatch resolution reaction. Instead of T4 DNA ligase, Taq DNA ligase was used. Pfu DNA polymerase (Stratagene) was employed in a parallel comparison to a set of reactions 15 that contained T4 DNA polymerase as the 3' exonuclease/polymerase. Reactions were carried out in Taq DNA ligase buffer containing 8 units of Taq DNA ligase (NEB), 2.5 units Taq DNA polymerase, 0.5 mM of each dNTP, various dilutions of CEL I, and either T4 DNA 20 polymerase or Pfu DNA polymerase). Reactions were run on a 1.5% TBE-agarose gels in the presence of ethidium bromide. It was found that in the presence of the Pfu DNA polymerase, Taq DNA polymerase, and Taq DNA ligase, the full-length integrity of the CEL I-treated substrate 25 DNA was enhanced compared to DNA incubated with CEL I alone. This result shows that enzymes with functionally equivalent activities can be successfully substituted into the GRAMMR reaction.

Example 3

**Restoration of Restriction Sites
to GFP Heteroduplex DNA after DNA Mismatch Resolution
(GRAMMR)**

This experiment teaches the operability of genetic reassortment by DNA mismatch resolution (GRAMMR) by demonstrating the restoration of restriction sites.

The full-length products of a twenty-fold scale-up of the GRAMMR reaction, performed at 37°C for one hour, using the optimal conditions found above (the 1x reaction contained sixty nanograms of heteroduplex DNA, one microliter of CEL I fraction five (described in Example 1), one unit T4 DNA polymerase in the presence of 2.5 units of Taq DNA polymerase and 0.2 units of T4 DNA ligase in 1 X NEB T4 DNA ligase buffer containing 0.5 mM of each dNTP in a reaction volume of 10 microliters) were gel-isolated and subjected to restriction analysis by endonucleases whose recognition sites overlap with mismatches in the GFP heteroduplex, thereby rendering those sites in the DNA resistant to restriction enzyme cleavage. The enzymes used were *Bam*HI, *Hind*III, *Hpa*I, and *Xho*I. Negative controls consisted of untreated GFP heteroduplex. Positive controls consisted of Cycle 3 or wild type GFP sequences, individually. All controls were digested with the same enzymes as the product of the DNA mismatch resolution reaction. All samples were run on a 2% TBE-agarose gel in the presence of ethidium bromide.

After treatment with the mismatch resolution cocktail, a proportion of the DNA gained sensitivity to *Bam*HI and *Xho*I restriction endonucleases, indicating that DNA mismatch resolution had occurred. The *Hpa*I-cut 5 samples could not be interpreted since a low level of cleavage occurred in the negative control. The *Hind*III, *Bam*HI and *Xho*I sites displayed different degrees of cleavage in the GRAMMR-treated samples. Restoration of the *Xho*I site was more extensive than that of the *Bam*HI 10 site, which was in turn, more extensive than restoration at *Hind*III site.

The extent to which cleavage occurs is indicative of the extent to which mismatches in the DNA 15 have been resolved at that site. Differences in mismatch resolution efficiency may relate to the nature or density of mismatches present at those sites. For example, the *Xho*I site spans a three-mismatch cluster, whereas the *Bam*HI site spans two mismatches and the *Hind*III site 20 spans a single mismatch.

Example 4

GRAMMR-Reassorted GFP Genes

25 This example demonstrates that GRAMMR can reassort sequence variation between two gene sequences in a heteroduplex and that there are no significant differences in GRAMMR products that were directly cloned, 30 or PCR amplified prior to cloning.

The GRAMMR-treated DNA molecules of Example 3 were subsequently either directly cloned by ligation into pCR-Blunt II-TOPO (Invitrogen), or amplified by PCR and ligated into pCR-Blunt II-TOPO according to the 5 manufacturer's instructions, followed by transformation into *E. coli*. After picking individual colonies and growing in liquid culture, DNA was prepared and the sequences of the GFP inserts were determined. As negative controls, the untreated GFP heteroduplex 10 substrate was either directly cloned or PCR amplified prior to cloning into the plasmid.

In GRAMMR, reassortment of sequence information results from a process of information transfer from one 15 strand to the other. These sites of information transfer are analogous to crossover events that occur in recombination-based DNA shuffling methods. For the purposes of relating the results of these reassortment experiments, however, the GRAMMR output sequences are 20 described in terms of crossovers. Sequences of twenty full-length GFP clones that were derived from the GRAMMR-treated GFP genes were analyzed. Four of these clones were derived from DNA that had been directly cloned into pZeroBlunt [ref] following GRAMMR treatment (no PCR 25 amplification). The other sixteen sequences were cloned after PCR amplification. Analysis of these full-length GFP sequences revealed that all twenty sequences had undergone sequence reassortment having between one and ten crossovers per gene. A total of 99 crossovers were 30 found in this set of genes, giving an average of about 5 crossovers per gene. With the distance between the first

and last mismatches of about 590 nucleotides, an overall frequency of roughly one crossover per 120 base-pairs was calculated. Within this set of twenty clones, a total of seven point mutations had occurred within the sequences 5 situated between the PCR primer sequences, yielding a mutation frequency of roughly 0.05%.

Thirty-five clones that had not been subjected to GRAMMR treatment were sequenced. Of these controls, 10 fourteen were derived from direct cloning and twenty-one were obtained after PCR amplification using the GFP heteroduplex as template. Of these thirty-five non-GRAMMR treated control clones, eight were recombinants, ranging from one to three crossovers, with most being 15 single crossover events. A total of twenty-five point mutations had occurred within the sequences situated between the PCR primers, yielding a mutation frequency of roughly 0.1%.

20 No significant differences were observed between the GRAMMR-treated products that were either directly cloned or PCR amplified. Notably, though, in the non-GRAMMR-treated controls, the frequency of recombinants was higher in the PCR amplified DNAs than in 25 the directly cloned DNAs. This higher frequency is consistent with results obtained by others in which a certain level of recombination was found to be caused by "jumping PCR." [Paabo, *et al.*, DNA damage promotes jumping between templates during enzymatic amplification. 30 J Biol Chem 265(90) 4718-4721].

Example 5

Heteroduplex Substrate Preparation for Plasmid-on- Plasmid Genetic Reassortment

5 **By DNA Mismatch Resolution (POP GRAMMR)**
of GFP Plasmids

This example teaches that heteroduplex substrate for Genetic Reassortment by DNA Mismatch

10 Resolution can be in the form of intact circular plasmids. Cycle 3-GFP and wild-type GFP heteroduplex molecules were prepared plasmid-on-plasmid (POP) format. In this format, the GFP sequences were reassorted within the context of a circular double-stranded plasmid vector

15 backbone. This made possible the recovery of the reassorted product by direct transformation of *E. coli* using an aliquot of the GRAMMR reaction. Consequently, neither PCR amplification nor other additional manipulation of the GRAMMR-treated DNA was necessary to

20 obtain reassorted clones.

Mismatched DNA substrate for POP-GRAMMR reactions was generated containing wild-type GFP (SEQ ID NO:01) and Cycle 3 GFP (SEQ ID NO:02), resulting in the 25 two pBluescript-based plasmids, pBSWTGFP (SEQ ID NO:03) and pBSC3GFP (SEQ ID NO:04), respectively. The GFPs were inserted between the *Kpn*I and *Eco*RI sites of the pBluescript polylinker so that the only sequence differences between the two plasmids occurred at sites 30 where the wild-type and Cycle 3 GFPs differ from one-another. Both plasmids were linearized by digestion of

the plasmid backbone with *SapI*, cleaned up using a DNA spin-column, mixed, amended to 1X PCR buffer (Barnes, 1994; *PNAS*, **91**, 2216-2220), heated in a boiling water bath for three minutes, and slow-cooled to room 5 temperature to anneal the denatured DNA strands.

Denaturing and annealing these DNAs led to a mixture of duplexes, the re-formation of parental duplexes, and the formation of heteroduplexes from the annealing of strands from each of the two input plasmids. Parental duplexes 10 were deemed undesirable for GRAMMR and were removed by digestion with restriction enzymes that cut in one or the other parental duplex but not in the heteroduplexed molecules. *PmlI* and *XhoI* were chosen for this operation since *PmlI* cuts only in the wild-type GFP sequence and 15 *XhoI* cuts only Cycle 3 GFP. After treatment with these enzymes, the products were resolved on an agarose gel. The full-length, uncut heteroduplex molecules were resolved from the *PmlI*- and *XhoI*-cut parental homoduplexes in an agarose gel and purified by excision 20 of the band and purification with a DNA spin column.

The resulting population of heteroduplexed molecules was treated with DNA ligase to convert the linear DNA into circular, double-stranded DNA 25 heteroduplexes. After confirmation by agarose gel-shift analysis, the circular double-stranded GFP heteroduplexed plasmid was used as substrate for GRAMMR reactions. Examples of the resulting clones are included as SEQ ID NO:05, SEQ ID NO:06, SEQ ID NO:07, and SEQ ID NO:08.

Example 6

**Exemplary Reaction Parameters
for Genetic Reassortment by DNA Mismatch Resolution**

5

CEL I and T4 DNA Polymerase Concentrations Compared

The GRAMMR reaction involves the interaction of numerous enzymatic activities. Several parameters associated with 10 the GRAMMR reaction were examined, such as CEL I concentration, T4 DNA polymerase concentration, reaction temperature, substitution of T4 DNA polymerase with T7 DNA polymerase, the presence of Taq DNA polymerase, and the source of the CEL I enzyme. A matrix of three 15 different CEL I concentrations versus two concentrations of T4 DNA polymerase was set up to examine the limits of the *in vitro* DNA mismatch resolution reaction.

Twenty-one nanograms (21ng) of the circular 20 double-stranded heteroduplexed plasmid, prepared as described above, was used as substrate in a series of ten microliter reactions containing 1X NEB ligase buffer, 0.5 mM each dNTP, 1.0 unit Taq DNA polymerase, 0.2 units T4 DNA ligase (Gibco/BRL), either 1.0 or 0.2 units T4 DNA 25 polymerase, and either 0.3, 0.1, or 0.03 microliters of a CEL I preparation (fraction 5, described in Example 1). Six reactions representing all six combinations of the 30 two T4 DNA polymerase concentrations with the three CEL I concentrations were prepared, split into equivalent sets of five microliters, and incubated at either 20 degrees C or 37 degrees C. A control reaction containing no CEL I

and 0.2 unit of T4 DNA polymerase with the other reaction components was prepared and incubated at 37 degrees C. After 30 minutes, one microliter aliquots of each reaction were transformed into competent DH5-alpha *E.* coli which were then plated on LB amp plates. Colonies were picked and cultured. Plasmid DNA was extracted and examined by restriction fragment length polymorphism analysis (RFLP) followed by sequence analysis of the GFP gene sequences. RFLP analysis was based on differences in several restriction enzyme recognition sites between the wild-type and Cycle 3 GFP genes. The RFLP results showed that throughout the CEL I /T4 DNA polymerase/temperature matrix, reassortment of restriction sites, that is GRAMMR, had occurred, and that no such reassortment had occurred in the zero CEL I control clones. DNA sequence analysis confirmed that reassortment had occurred in all of the CEL I -containing samples. Sequencing also confirmed that the zero- CEL I controls were not reassorted, with the exception of a single clone of the 16 control clones, which had a single-base change from one gene sequence to the other, presumably resulting either from repair in *E. coli* or from random mutation. The sequences of several exemplary GRAMMR-reassorted GFP clones are shown; all of which came from the reaction containing 0.3 microliters of the CEL I preparation and 1.0 unit of T4 DNA polymerase incubated at 37 degrees C. The parental wild-type and Cycle 3 GFP genes are shown first for reference.

Example 7**Taq DNA Polymerase is Not Required
for Genetic Reassortment by DNA Mismatch Resolution**

5

This experiment teaches that Taq DNA Polymerase does not dramatically, if at all, contribute or interfere with the functioning of Genetic Reassortment by DNA Mismatch Resolution (GRAMMR). Taq DNA polymerase is reported to have a 5' flap-ase activity, and had been included in the teachings of the previous examples as a safeguard against the possible formation and persistence of undesirable 5' flaps in the heteroduplexed DNA undergoing GRAMMR.

10

15

GRAMMR reactions were set up, as in Example 6, with twenty-one nanograms of the circular double-stranded heteroduplexed GFP plasmid substrate in ten microliter reactions containing 1X NEB ligase buffer, 0.5 mM each dNTP, 0.2 units T4 DNA ligase, 1.0 unit T4 DNA polymerase, 1.0 microliter of a CEL I preparation (fraction 5, described in Example 1), and either 2.5 units, 0.5 units of Taq DNA polymerase, or no Taq DNA polymerase. After 30 minutes, one microliter aliquots of each reaction were transformed into competent DH5-alpha *E. coli* which were then plated on LB amp plates.

Colonies were picked and cultured. Plasmid DNA was extracted and examined by RFLP analysis followed by sequence analysis of the GFP gene sequences. The RFLP results showed that reassortment of restriction sites, that is, GRAMMR, had occurred both in the presence and

PUBLISHED BY
THE GENETIC ENGINEERING
AND BIOTECHNOLOGY
COMMISSION
GOVERNMENT OF INDIA

the absence of Taq DNA polymerase in the GRAMMR reaction. DNA sequence analysis confirmed these results. Therefore, the data shows that Taq DNA polymerase was unnecessary for GRAMMR.

5

Example 8

Alternate Proofreading DNA Polymerases for Genetic Reassortment by DNA Mismatch Resolution

10

This experiment teaches that Genetic Reassortment by DNA Mismatch Resolution is not limited to the use of T4 DNA polymerase, and that alternate DNA polymerases can be substituted for it.

15

Reactions were set up, as in Example 6, with twenty-one nanograms of the circular double-stranded heteroduplexed GFP plasmid substrate in ten microliter reactions containing 1X NEB ligase buffer, 0.5 mM each 20 dNTP, 0.2 units T4 DNA ligase (Gibco/BRL), 10 units or 2 units of T7 DNA polymerase, 1.0 microliter of a CEL I preparation (fraction 5, described in Example 1), and 2.5 units of Taq DNA polymerase. After 30 minutes, one microliter aliquots of each reaction were transformed 25 into competent DH5-alpha *E. coli* which were then plated on LB amp plates. Colonies were picked and cultured. Plasmid DNA was extracted and examined by RFLP analysis followed by sequence analysis of the GFP gene sequences. The RFLP results showed that reassortment of restriction 30 sites, that is GRAMMR, had occurred in both T7 DNA polymerase-containing reactions. DNA sequence analysis

confirmed these results. Therefore, the data shows that T7 DNA polymerase can substitute for T4 DNA polymerase for GRAMMR. In addition, it shows that individual components and functionalities can be broadly substituted 5 in GRAMMR, while still obtaining similar results.

Example 9

Use of Cloned CEL I in the GRAMMR Reaction

10 This example teaches that CEL I from a cloned source can be used in place of native CEL I enzyme purified from celery in Genetic Reassortment By DNA Mismatch Resolution without any noticeable change in results.

15 The cDNA of CEL I was cloned from celery RNA. The gene was inserted into a TMV viral vector and expressed. Transcripts of the construct were used to infect *Nicotiana benthamiana* plants. Infected tissue was 20 harvested, and the CEL I enzyme was purified. The GRAMMR results obtained using the purified enzyme were compared to those using CEL I purified from celery, and were found to be similar.

25 Reactions were set up using twenty-one nanograms of the circular double-stranded heteroduplexed GFP plasmid substrate in ten microliters containing 1X NEB ligase buffer, 0.5 mM each dNTP, 0.2 units T4 DNA ligase (Gibco/BRL), 1 unit of T4 DNA polymerase, and 30 either 1.0 microliter of CEL I purified from celery (fraction 5, described in Example 1), or 0.3 microliters

of CEL I purified from a cloned source. After 30 minutes, one microliter aliquots of each reaction were transformed into competent DH5-alpha *E. coli* which were then plated on LB amp plates. Colonies were picked and cultured.

5 Plasmid DNA was extracted and examined by RFLP analysis followed by sequence analysis of the GFP gene sequences. The RFLP results showed that reassortment of restriction sites, that is, GRAMMR had occurred in both celery-derived CEL I, as well as cloned CEL I-containing

10 reactions. DNA sequence analysis confirmed these results. Therefore, the data shows CEL I from a cloned source can be used in lieu of CEL I from celery for GRAMMR. In addition, the data demonstrates that it is CEL I activity that is part of the GRAMMR method, rather

15 than a coincidental effect resulting from the purifying steps used in extracting CEL I from celery.

Example 10

20 **Molecular Breeding of
Tobamovirus 30K Genes in a Viral Vector.**

In the preceding examples, Genetic Reassortment by DNA Mismatch Resolution has been taught to be useful
25 for reassorting sequences that are highly homologous, for example, wtGFP and Cycle 3 GFP are 96% identical. The present example teaches that GRAMMR can be used to reassort more divergent nucleic acid sequences, such as genes encoding tobamovirus movement protein genes.

30 Heteroduplexes of two tobamovirus movement protein (MP) genes that are approximately 75% identical

were generated. The heteroduplex substrate was prepared by annealing partially-complementary single-stranded DNAs of opposite strandedness synthesized by asymmetric PCR; one strand encoding the movement protein gene from the 5 tobacco mosaic virus U1 type strain (TMV-U1) (SEQ ID NO:09), and the other strand encoding the movement protein gene from tomato mosaic virus (ToMV) (SEQ ID NO:10). The sequences of the two partially complementary movement protein genes were flanked by 33 nucleotides of 10 absolute complementarity to promote annealing of the DNAs at their termini and to facilitate PCR amplification and cloning. The annealing reaction took place by mixing 2.5 micrograms of each single-stranded DNA in a 150 microliter reaction containing 333 mM NaCl, 33 mM MgCl₂, 15 3.3 mM dithiothreitol, 166mM Tris-HCl, pH 7, and incubating at 95°C for one minute followed by slow cooling to room-temperature. GRAMMR was performed by incubating 5 microliters of the heteroduplex substrate in a 20 microliter reaction containing 1X NEB ligase buffer, 20 0.5 mM each dNTP, 0.4 units T4 DNA ligase (Gibco/BRL), 2.0 units of T4 DNA polymerase, and CEL I. The CEL I was from a cloned preparation and the amount that was used varied from 2 microliters of the prep, followed by five serial 3-fold dilutions. A seventh preparation with no 25 CEL I was prepared, which served as a control.

After one hour at room-temperature, DNA was purified from the reactions using Strataprep spin DNA purification columns (Stratagene, LaJolla, CA) and used 30 as templates for PCR reactions using primers designed to anneal to the flanking primer-binding sites of the two

sequences. PCR products from each reaction were purified using Strataprep columns, digested with *AvrII* and *PacI*, and ligated into the movement protein slot of similarly-cut pGENEWARE-MP-*Avr-Pac*. This plasmid contained a full-length infectious tobamovirus-GFP clone modified with *AvrII* and *PacI* sites flanking the movement protein gene to permit its replacement by other movement protein genes. After transformation of DH5-alpha *E. coli* and plating, colonies were picked, cultures grown, and DNA was extracted. The movement protein inserts were subjected to DNA sequence analysis from both directions and the sequence data confirmed that in the majority of inserts derived from the GRAMMR-treated material were reassorted sequences made up of both TMV-U1 and ToMV movement protein gene sequences. The DNA sequences of several exemplary GRAMMR MP clones are shown as SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, and ID NO:15.

20

EXAMPLE 11**GRAMMR Reassortment to Generate Improved Arsenate Detoxifying Bacteria**

25 Arsenic detoxification is important for mining of arsenopyrite-containing gold ores and other uses, such as environmental remediation. Plasmid pGJ103, containing an arsenate detoxification operon (Ji and Silver, 1992) (Ji, G. and Silver, S., Regulation and expression of the arsenic resistance operon from *Staphylococcus aureus* plasmid pI258, *J. Bacteriol.* 174, 3684-3694 (1992), incorporated herein by reference), is obtained from Prof.

Simon Silver (U. of Illinois, Chicago, Ill.). *E. coli* TG1 containing pGJ103, containing the pI258 *ars* operon cloned into pUC19, has a MIC (minimum inhibitory concentration) of 4 μ g/ml on LB ampicillin agar plates.

- 5 The *ars* operon is amplified by mutagenic PCR [REF], cloned into pUC19, and transformed into *E. coli* TG1. Transformed cells are plated on a range of sodium arsenate concentrations (2, 4, 8, 16 mM). Colonies from the plates with the highest arsenate levels are picked.
- 10 The colonies are grown in a mixed culture with appropriate arsenate selection. Plasmid DNA is isolated from the culture. The plasmid DNA is linearized by digestion with a restriction endonuclease that cuts once into the pUC19 plasmid backbone. The linearized plasmids
- 15 are denatured by heating 10 min. at 94°C. The reaction is allowed to cool to promote annealing of the single strands. Partially complementary strands that hybridize have non-basepaired nucleotides at the sites of the mismatches. Treatment with CEL I (purified by the method
- 20 of Example 9) causes nicking of one or the other polynucleotide strand 3' of each mismatch. The presence of a polymerase containing a 3'-to-5' exonuclease ("proofreading") activity, such as T4 DNA polymerase allows excision of the mismatch, and subsequent 5'-to-3'
- 25 polymerase activity fills in the gap using the other strand as a template. T4 DNA ligase then seals the nick by restoring the phosphate backbone of the repaired strand. The result is a randomization of mutations among input strands to give output strands with potentially
- 30 improved properties. These output polynucleotides are transformed directly into *E. coli* TG1 and the cells are

plated at higher arsenate levels; 8, 16, 32, 64 mM. Colonies are picked from the plates with the highest arsenate levels and another round of reassortment is performed as above except that resulting transformed 5 cells are plated at 32, 64, 128, 256 mM arsenate. The process can then be repeated one or more times with the selected clones in an attempt to obtain additional improvements.